CERTIFICATE OF ELECTRONIC FILING
I hereby certify that this correspondence is being electronically filed
With the U.S. Patent & Trademark Office on 21 February 2007..

/Lynne M. Milliot/
Lynne M. Milliot

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	Group Art Unit: 2179
Mark W. DAVIS et al.	
Application No. 09/893,299	Examiner: Joshua D. CAMPBELL
Confirmation No. 1994	
Filed: 27 June 2001	**CUSTOMER NO. 22470**
Title: **Method and Apparatus for Duplicate Detection**	

MAIL STOP APPEAL BRIEF - PATENTS
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

AMENDED APPEAL REPLY BRIEF

Sir:

This Amended Appeal Reply Brief is filed in response to the Examiner's Answer mailed 21 December 2006, and in further support of Appellants' appeal from the Final Office Action, mailed 16 February 2005 in this case. A Request for Oral Hearing accompanies this submission and the required fee of $1,000 is being electronically submitted herewith.

Should it be determined that any additional fees are required in connection with this communication, the Commissioner is hereby authorized to charge those fees to Deposit Account No. 50-0869 (Attorney Docket No. INXT 1016-1).

TABLE OF CONTENTS

I. AGREED POINTS

The Examiner's Answer (hereafter "EA") acknowledges Appellants' statement identifying the real party in interest, the lack of related appeals or interferences, the status of amendments after final, the summary of invention and the issues on appeal. It also acknowledges the copy of appealed claims in the appendix and takes no issue with the evidentiary appendix.

II. RELATED APPEALS AND INTERFERENCES

There are no known appeals or interferences relating to this case.

III. STATUS OF CLAIMS

The EA notes a typographical error in the opening brief. The Examiner is correct that claims 1-12 are pending in this case and are subject to this appeal.

IV. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

The EA insists that Prager is the base reference and that Pugh be used to modify Prager. This puts the Examiner in an impossible position, using a primary reference that has no mention of the function, way or result corresponding to the claims on appeal. Given this written insistence (EA 3, § 6), Appellants accept that the Examiner will not argue for Pugh as a base reference and reply accordingly.

V. REPLY TO EXAMINER'S ARGUMENTS

Rejection of claims 1-12 under 35 U.S.C. 103(a) as being unpatentable over Prager in view of Pugh should be reversed.

A. Preliminary Review of the Technology Disclosed and References

1. The Disclosed Technology

The Opening Brief explained the use of triangulation for duplicate detection in a system that does not generate fingerprints or similar hashes of documents. (OB 2-3) The Examiner says that he does not see triangulation in claims 1 and 11. (EA 6-7) Therefore, we revisit and expand our explanation of so-called triangulation.

Claim 1 reads:

> *A method of detecting duplicates in a set of documents having associated nearest neighbor similarity scores, the method including:*

> *for a particular document in the set of documents, selecting nearest neighbors of the particular document; and*
>
> *flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are identical.*

We explained that, "As a convenient shorthand, we will follow the terminology used in the specification and refer to detecting duplicates A & B based on their similarity to C as 'triangulation.'" (OB 1)

In this reply, we tie A, B and C to the words of claim 1, to illustrate triangulation. Let A be a candidate document being classified using K nearest neighbor (KNN) or similar technology. In the words of the claim, A will be the *particular document*. From the preamble, this *particular document* belongs to *the set of documents* and has a list of *associated nearest neighbor similarity scores*. Let B_1, B_2, B_3, etc. be other candidate documents in *the set of documents* that qualify to be *selected* as *nearest neighbors of the particular document*. We know from the claim preamble that all of the documents A, B_1, B_2, B_3, have *associated* lists of *nearest neighbor similarity scores*, for instance, from being compared, for instance, to training set documents C_1, C_2, C_3. (As explained below, Prager introduces a centroid method in which documents **C** are actually composites representing categories, rather than individual documents belonging to categories.) Call the resulting similarity scores AC_1, AC_2, AC_3, B_1C_1, B_1C_2, B_1C_3, B_2C_1, etc. In the flagging step, the list of *similarity scores* associated with document A is compared to similar lists for its nearest neighbors B_1, B_2, B_3. For documents A and B_1, for instance, the *similarity scores* AC_1, AC_2, AC_3 would be compared to B_1C_1, B_1C_2, B_1C_3. If the lists of similarity scores match, documents A and B_1 are flagged as potential duplicates. We call this triangulation, because it involves comparison of similarity scores for A and B_1 against **C**, instead of direct comparison between the term vectors for A and B_1.

Triangulation among the documents A, B and C can similarly be tied to the words of claim 11.

2. The Prager Reference

The EA makes very limited reference to the primary reference Prager, focusing on col. 1 line 55 – col. 2 line 42, which is the background discussion that precedes Prager's invention disclosure. (EA 3-4) Further reference to col. 4 line 34 – col. 5 line 3

(EA 4-5) relates only to dependent claims 5-9, which stand or fall with independent claim 1.

Two sorts of similarity scores are described in Prager's background section. First, centroid technology calculates measures of similarity of term vectors that represent documents A and B with term vectors that represent multiple categories X. Prager, col. 1 line 56 – col. 2 line 16. Category term vectors **X** are constructed as aggregates of documents. Second, KNN technology calculates a measure of similarity of term vectors for documents A and B with term vectors for multiple documents C in a training or reference set. Prager, col. 2 lines 17-42. Applying centroid technology, term vectors for candidate documents A and B are compared to *m* vectors **X**, where *m* is the number of categories into which the document might be assigned. The closest score wins. Applying KNN technology, the term vectors for candidate documents A and B are compared to *n* vectors **C**, where *n* is the number of documents in the training set. Typically, a training set has many examples of documents C in each category X, so n > m.

Prager's invention, to which the EA makes no reference or citation, presents an enhancement to the centroid method. By this enhancement, Prager helps determine when a new category should be created that includes features of two existing categories.

The background section of the primary Prager reference does not present any function, way or result that corresponds to the claimed method of duplicate detection. Duplicates are not mentioned anywhere in Prager.

3. The Pugh Reference

The secondary reference Pugh does not use similarity scores, in the sense of calculating metrics of similarity between documents. The word "score" does not appear anywhere in Pugh.

Near duplicates are determined by counting the number of "fingerprints" that exactly match between two documents. Reading Pugh carefully, document fingerprints, like human fingerprints, either match or do not match. In column 9, lines 40-42, Pugh writes, "In one embodiment of the invention, if two documents have any one fingerprint in common, they are considered to be 'near-duplicates'." In more detail, Pugh § 4.3.2.4 explains figure 7, "[A]s indicated by the loop 740-760 through each fingerprint of the

second document, nested within the loop 730-770 through each fingerprint of the first document, the **fingerprints are compared to determine whether or not they match**". Col. 14 lines 12-16. "[O]nly documents with **matching fingerprints** need be analyzed." Col. 14 lines 41-42. Therefore, fingerprints do not indicate degrees of similarity, they only match or fail to match.

The number of fingerprints that Pugh generates for a document depends on a hashing function that assigns document features to one of a predetermined number of lists. Col. 12. Hash functions that generated three to nine lists of document features for fingerprinting were tested by Pugh. Col. 9 lines 4-8. These hash functions assigned extracts from a document to multiple lists from which multiple fingerprints were generated.

Using a hashing function to generate multiple lists for fingerprinting is the essential improvement that Pugh emphasizes, by comparison to prior implementations of Broder-Rabin fingerprinting. *See* col. 3 lines 24-57 (describing computational problem with fingerprinting of paragraphs, sentences, words or shingles), col. 10 lines 12-22 (suggesting pre-filtering step to generate reduced number of fingerprints) & col. 13 lines 51-65 (referencing Broder and Rabin articles). One of skill in the art would understand Pugh as emphasizing a new method for generating a fixed number of fingerprints from a document of arbitrary size.

To avoid potential confusion, we note that Pugh uses two hashes, one to set up lists and another to generate fingerprints from lists. Pugh builds on the Broder-Rabin fingerprinting hashes by adding a list-generating hash for generating a fixed number of lists for fingerprinting from a document of arbitrary size.

Despite carefully reading Pugh, we are at a complete loss to see what the EA considers to be "detection scores". (EA 4, line 3 & *passim*) There is no citation to Pugh that would identify what the "detection scores" are supposed to be. The only thing close to a score that we see is a count of the number of fingerprints that match between two documents. (EA 4, lines 3-5) The cited passage, Pugh cols. 7-8, is the wrong part of the disclosure to read, if one is trying to learn Pugh's method; the cited passage does not refer to detection scores of any type, only to a transitivity principle that is not relevant to the claims or to this appeal.

4. Trying to Combine Prager and Pugh

The EA does not explain how to combine the references. (EA 3-4) We noted this in our opening brief (OB 4, § VII.A.2) but failed to provoke a response. That is, the Examiner continues to leave us wondering wonder what parts of Pugh the Examiner would add to Prager and whether Pugh would have to be modified after being added to Prager.

The motivation to eliminate duplicates from a document set is not remarkable. (EA 7-8) It is a long-felt need, as Pugh explains, that has been addressed in various ways over the years. But the motivation to eliminate duplicates does not explain how one of ordinary skill would combine Pugh with Prager.

The clearest way to combine Pugh with Prager would be to use two hashing algorithms, one to generate multiple term vectors and the other to generate multiple fingerprints. The multiple term vectors would double for generating multiple or intermediate similarity scores and for generating multiple fingerprints. Pugh emphatically teaches using two hashes. Pugh's essential teaching is using the second hash to create multiple extract lists, which are analogous to multiple term vectors. Multiple extract lists lead to a determinate and manageably small number of fingerprints to be compared between documents. In this combination, Pugh added to Prager would enable counting the number of matching fingerprints to detect near-duplicates.

There is no teaching or suggestion to combine Pugh with Prager, then modify Pugh to use Prager's similarity scores instead of fingerprints. Prager does not mention fingerprinting and Pugh does not mention similarity scores. To combine Pugh's method with Prager logically would involve generating fingerprints, because similarity scores do not have the unique characteristics of fingerprints that Pugh emphasizes. *See* Pugh, § 4.3.2.3, especially col. 13 lines 52-53 (collision-free or low probability of collision). The most likely combination between Prager and Pugh's respective disclosures would be applying Pugh's multiple list generation to generate multiple term vectors, so that both fingerprints and similarity scores could be generated from the same multiple term vectors (instead of one term vector per document.) Both similarity scores and fingerprints would still need to be generated, because they have different properties. Duplicates might be eliminated before similarity scores were calculated.

Nothing in either reference suggests using Prager's similarity scores in place of Pugh's fingerprints.

There is no teaching or suggestion to modify Pugh to triangulate documents A and B with document set **C**. That is, because comparing fingerprints is a shorthand for comparing document A vs. document B, there is no sense in which Pugh teaches comparing document A to **C** and document B to **C** and then comparing the similarity scores AC_1, AC_2, AC_3 with B_1C_1, B_1C_2, B_1C_3 to decide whether A and B are potential duplicates.

With this understanding of the disclosed and referenced technology in mind, we turn to the arguments on appeal.

B. Rejection of claims 1 and 11 under Section 103(a) as being unpatentable over Prager in view of Pugh is improper

1. Neither Prager nor Pugh teaches triangulation

Our position on appeal (OB 2-4) has been that rejection of claims 1 and 11 is improper because triangulation to detect duplicates using nearest neighbor similarity scores is not taught by either reference or by the combination.

The EA counters that the Examiner does not see triangulation in the claims. (EA 6, "it is unclear to the examiner where the 'triangulation element' exists in claims 1 and 11") Often, differences between an Examiner and Applicants are a matter of misunderstanding. That seems to be the case here, so we elaborate on our explanation.

We explain above, *supra* at 1, how triangulation is expressed in the words of the claim. In the flagging element, finding identical lists or sublists of nearest neighbor similarity scores necessarily involves triangulation. Suppose you are investigating duplication between documents A and B, using the claimed method. It makes no sense to compare the similarity scores AB and BA, because they would always be exactly the same. So comparing similarity scores AC_1, AC_2, AC_3 with B_1C_1, B_1C_2, B_1C_3 makes flagging workable. This is what we call triangulation.

Pugh does not teach triangulation. For documents A and B, Pugh generates fingerprints A_1, A_2, A_3 and B_1, B_2, B_3. Then, Pugh compares the respective fingerprints for A and B. If enough fingerprints match exactly, the documents are considered

potential duplicates. Pugh emphasizes use of hashes that produce unique fingerprints; there is no sense in fingerprint comparison of similarity in a metric or distance sense, only that fingerprints match or do not.

The Examiner cites Pugh col. 7-8, which is a discussion of transitivity. (EA 7) Transitivity is that mathematical or logical property that A = B and B = C implies A = C. In the col. 7-8 passage, Pugh fingerprinting suggests that when one has (a) one cluster of multiple documents that are near duplicates, based on multiple identical fingerprints, and (b) another cluster of documents that are near duplicates, based on multiple identical fingerprints, and (c) one document (completely self-identical) that belongs to both clusters, then the clusters can be combined and treated as one. It is tenuous, at best, to extend transitivity from fingerprints in Pugh to Prager's method. There is no teaching, suggestion or evidence in Prager that transitivity would hold true for near duplication, as opposed to exact duplication. Moreover, the proposed transitivity between clusters of near duplicates does not read on claim 1 or 11, because it does not involve comparing similarity scores AC_1, AC_2, AC_3 with B_1C_1, B_1C_2, B_1C_3.

Neither the primary reference Prager, which teaches an improved centroid technology, nor secondary reference Pugh, which teaches improved fingerprinting, teaches triangulation. There is no teaching or suggestion in either reference or argued on appeal (beyond simply the motivation to eliminate duplicates) that would lead one of ordinary skill to combine the references to invent triangulation for duplicate detection.

Therefore, the rejections should be reversed.

2. Combining Pugh with Prager to include triangulation is inventive

In the absence of a teaching or suggestion, explicit or implicit, to combine the elements of Pugh and Prager in the manner claimed, taken as a whole, the combination should be considered inventive.

We return to our argument that there is no evidentiary quality suggestion to combine the references **in the manner claimed**. (OB 4-6) The Examiner cites limited evidence (EA 7) from cols. 7 – 8, where transitivity is discussed. This is the wrong passage to cite for the proposition that elimination of duplicates is desirable. *Compare* EA 7-8 with Pugh col. 3. This passage does not teach or suggest arranging the elements claimed in the manner claimed, taken as a whole. One of ordinary skill in the

art would not understand Pugh's proposed transitivity between clusters of near duplicate documents to teach substitution of similarity scores and triangulation as claimed on appeal, for the two hashes and unique fingerprints as taught by Pugh.

The law is clear and not disputed that combining the references using the claim as a blueprint (20-20 hindsight) is impermissible. 2-5 Chisum on Patents § 5.03 [2][c] n. 29 (2005 Lexis version); *e.g. ATD Corp. v. Lydall, Inc.*, 159 F.3d 534, 546, 48 USPQ2d 1321, 1329 (Fed. Cir. 1998) ("Determination of obviousness can not be based on the hindsight combination of components selectively culled from the prior art to fit the parameters of the patented invention."); *Grain Processing Corp. v. American Maize-Products Corp.*, 840 F.2d 902, 907, 5 USPQ2d 1788, 1792 (Fed. Cir. 1988) ("Care must be taken to avoid hindsight reconstruction by using 'the patent in suit as a guide through the maze of prior art references, combining the right references in the right way so as to achieve the result of the claims in suit.' ").

The statutory "as a whole" rule of Section 103 prohibits this use of hindsight to combine elements in a manner not taught or suggested by the references. The Federal Circuit explained in *Ruiz v. A.B. Chance*, 357 F.3d 1270, 1275, 69 U.S.P.Q.2d (BNA) 1686 (Fed. Cir. 2004):

> In making the assessment of differences, section 103 specifically requires consideration of the claimed invention "as a whole." Inventions typically are new combinations of existing principles or features. *Envtl. Designs, Ltd. v. Union Oil Co.*, 713 F.2d 693, 698 (Fed. Cir. 1983) (noting that "virtually all [inventions] are combinations of old elements."). The "as a whole" instruction in title 35 prevents evaluation of the invention part by part. Without this important requirement, an obviousness assessment might break an invention into its component parts (A + B + C), then find a prior art reference containing A, another containing B, and another containing C, and on that basis alone declare the invention obvious. This form of hindsight reasoning, using the invention as a roadmap to find its prior art components, would discount the value of combining various existing features or principles in a new way to achieve a new result - often the very definition of invention.
>
> Section 103 precludes this hindsight discounting of the value of new combinations by requiring assessment of the invention as a whole. This court has provided further assurance of an "as a whole" assessment of the invention under § 103 by requiring a showing that an artisan of ordinary skill in the art at the time of invention, confronted by the same problems as the inventor and with no knowledge of the claimed invention, would select the various elements from the prior art and combine them in the claimed manner. In other words, the examiner or court must show some

> suggestion or motivation, before the invention itself, to make the new combination. *See In re Rouffet*, 149 F.3d 1350, 1355- 56 (Fed. Cir. 1998).

See, *Princeton Biochemicals, Inc. v. Beckman Coulter, Inc.*, 411 F.3d 1332, 1337, 75 U.S.P.Q.2d (BNA) 1051 (Fed. Cir. 2005) (reciting *Ruiz* rule; "simply identifying all of the elements in a claim in the prior art does not render a claim obvious"). The Federal Circuit has rejected the Examiner's approach of using a general motivation, like "eliminate duplicates", to combine elements in a manner, such as triangulation, that is not taught or suggested. This general approach has been rejected both as lacking the required evidentiary support (*In re Lee,* 277 F.3d 1338, 1342-44, 61 USPQ2d at 1433-34 (Fed. Cir. 2002); *In re Kotzab*, 217 F.3d 1365, 1369-70 (Fed. Cir. 2000); *Kolmes v. World Fibers Corp.*, 107 F.3d 1534, 1541 (Fed. Cir. 1997)) and because the logic applied is prohibited by statute (*Ruiz; Princeton Biochemicals*).

Therefore, rejection of claims 1 and 11 should be reversed.

C. Rejection of claims 3 and 4 under 35 U.S.C. 103(a) is improper

We separately argued claims 3 and 4 (OB 11), because the claims require that the nearest neighbor similarity scores [3] or the k nearest neighbors [4] have been calculated or determined prior to duplicate detection for a different purpose than duplicate detection. Foresight in retaining the similarity scores and nearest neighbor lists is rewarded with efficiency in duplicate elimination.

The response (EA 10) begins by arguing that Prager col. 2, lines 17-33 teaches using similarity scores for the purpose of categorization, which is true.

The argument (EA 10) goes flat with the mistaken argument that "Pugh teaches that these scores may then be used for potentially [sic] duplicate document detection (column 7, line 26-column 8, line 28 of Pugh)." This argument is unsupported by the passage cited. The transitivity passage has nothing to do with using similarity scores and triangulation in place of two hashes to produce a predetermined number of fingerprints. Pugh, for Google, would not agree with the Examiner's reading of transitivity as suggesting computation of a multitude of similarity scores in place of a few fingerprints. Regardless of how much latitude the Examiner has to interpret the intended purpose of Pugh, there is no dispute that Pugh wrote about and disclosed a two-hash method for reducing the number of fingerprints that would need to be calculated and compared to detect duplicates. With Pugh's whole disclosure in mind, there is no

possibility that Pugh intended to suggest or any passage could reasonably be considered to suggest decreasing efficiency by calculating similarity scores instead of double-hashing to produce a reduced number of fingerprints.

Moreover, it would make more sense in a combination to calculate Pugh's fingerprints first and eliminate duplicates, before the more expensive calculation of similarity scores and nearest neighbors. Nothing in Prager or Pugh teaches the order of calculations, as between duplicate elimination and categorization.

For these reasons, rejection of claims 3 and 4 should be reversed.

CONCLUSION

In view of the foregoing, Appellants ask that this honorable Board reverse the Examiner's rejections of the claims. In addition, it is submitted that all claims which are the subject of this examination are now allowable, and a notice of intent to issue a patent is respectfully requested.

Fee Authorization. The Commissioner is hereby authorized to charge any additional fee(s) determined to be due in connection with this communication, or credit any overpayment, to our Deposit Account No. 50-0869 (File No. INXT 1016-1).

Respectfully submitted,

Dated: 21 February 2007

/Ernest J. Beffel, Jr./
Ernest J. Beffel, Jr., Reg. No. 43,489
Attorney for Patent Owner

HAYNES BEFFEL & WOLFELD LLP
P.O. Box 366
751 Kelly Street
Half Moon Bay, CA 94019
Telephone: (650)712.0340
Facsimile: (650)712.0263

CLAIMS APPENDIX

1. (Original) A method of detecting duplicates in a set of documents having associated nearest neighbor similarity scores, the method including:

for a particular document in the set of documents, selecting nearest neighbors of the particular document; and

flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are identical.

2. (Original) The method of claim 1, further including flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are within a tolerance t of one another.

3. (Original) The method of claim 1, wherein the nearest neighbor similarity scores are calculated prior to duplicate detection for a different purpose than the duplicate detection and stored with the documents.

4. (Original) The method of claim 1, wherein the k nearest neighbors are determined prior to duplicate detection for a different purpose than the duplicate detection and stored with the documents.

5. (Original) The method of claim 1, wherein the documents are text documents.

6. (Original) The method of claim 5, wherein the text documents include visual formatting.

7. (Original) The method of claim 1, wherein the documents are voice recordings.

8. (Original) The method of claim 1, wherein the documents are musical performances.

9. (Original) The method of claim 1, wherein the documents are graphic images.

10. (Original) The method of claim 1, wherein the nearest neighbors are limited to k nearest neighbors.

11. (Original) A method of detecting duplicates in a set of documents, the method including:

identifying nearest neighbors of documents in the set of documents, based on nearest neighbor similarity scores;

for a particular document in the set of documents, flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are identical.

12. (Original) The method of claim 11, further including flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are within a tolerance t of one another.

EVIDENCE APPENDIX

Appellants have no evidence to submit under 37 CFR 1.130, 1.131 or 1.132.

RELATED PROCEEDINGS APPENDIX

As there are no related proceedings, there is nothing to submit in this appendix.